Annals of the Institute of Statistical Mathematics (2021) 73:519–533
https://doi.org/10.1007/s10463-020-00758-z

Instrument search in pseudo-likelihood approach for nonignorable nonresponse

Ji Chen[1] · **Jun Shao**[2,3] · **Fang Fang**[2]

Received: 7 May 2019 / Revised: 2 December 2019 / Published online: 13 June 2020

Abstract
With nonignorable nonresponse, an effective method to construct valid estimators of population parameters is to use a covariate vector called instrument that can be excluded from the nonresponse propensity, but are associated with the response even when other covariates are conditioned. The existing work in this approach assumes such an instrument is given, which is frequently not the case in applications. In this paper, we investigate how to search for an instrument from a given set of covariates, based on a pseudo likelihood approach assuming a parametric distribution of response conditioned on covariates and a totally unspecified nonresponse propensity. We propose a method and show that it produces a consistent instrument selection as the sample size tends to infinity, under some regularity conditions. The proposed method is examined in a simulation study and illustrated in a real data example.

Keywords Nonignorable nonresponse · Nonresponse instrument · Pseudo-likelihood · Variable selection

1 Introduction

Nonresponse with an appreciable rate is common in many applications such as clinical trials and sample surveys. Let Y be a response or outcome of interest that may have nonresponse, X be a covariate vector that is always observed, and R be the indicator equaling 1 if Y is observed and 0 if Y is missing. When the propensity $P(R = 1|Y, X)$ is equal to $P(R = 1|X)$ not depending on Y, the nonresponse is called

✉ Fang Fang
ffang@sfs.ecnu.edu.cn

[1] School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai 200241, China

[2] KLATASDS-MOE, School of Statistic, East China Normal University, 500 Dongchuan Road, Shanghai 200241, China

[3] Department of Statistics, University of Wisconsin-Madison, 1300 University Ave., Madison, WI 53706, USA

ignorable and there is a rich literature on methodology of handling ignorable nonresponse (Little and Rubin 2002). However, in many applications, $P(R = 1|Y, X)$ depends on both X and Y, in which cases nonresponse is referred to as nonignorable and estimation of population parameters is much more challenging than that in the case of ignorable nonresponse.

Throughout we use $p(\cdot|\cdot)$ or $p(\cdot)$ as a generic notation for the conditional or unconditional probability density with respect to an appropriate measure (discrete, continuous, or mixed). With nonignorable nonresponse, when both $p(Y|X)$ and $P(R = 1|Y, X)$ are parametric, maximum-likelihood methods have been developed (Greenlees et al. 1982; Baker and Laird 1988). When both $p(Y|X)$ and $P(R = 1|Y, X)$ are nonparametric, Robins and Ritov (1997) showed that the population may not be identifiable. Hence, efforts have been made to develop semiparametric methods, assuming one of $p(Y|X)$ and $P(R = 1|Y, X)$ has a parametric form and the other one is nonparametric. Qin et al. (2002) and Wang et al. (2014) imposed a parametric model on $P(R = 1|Y, X)$, but allowed $p(Y|X)$ to be nonparametric. Following Tang et al. (2003), Zhao and Shao (2015), and Chen et al. (2018), in this paper, we focus on a nonparametric $P(R = 1|Y, X)$ and a parametric model:

$$p(Y|X) = f(Y|X;\theta), \tag{1}$$

where θ is an unknown parameter vector and $f(Y|X;\theta)$ is known when θ is known.

In their semiparametric approach, Zhao and Shao (2015) utilized a covariate vector Z called nonresponse instrument or simply instrument, to guarantee the identifiability of population parameters, so that consistent estimators can be obtained. More precisely, an instrument Z is a sub-vector of X, i.e., $X = (U, Z)$, such that Z satisfies the following two conditions:

$$P(R = 1|Y, X) = P(R = 1|Y, U), \tag{2}$$

$$p(Y|X) = p(Y|U, Z) \text{ depends on } Z. \tag{3}$$

If we know which components of X satisfy (2)-(3), then parameters in $p(Y|X)$ can be estimated using pseudo-likelihoods (Zhao and Shao 2015) and parametric model selection regarding (1) can also be performed (Fang and Shao 2016). Note that even if both $p(Y|X)$ and $P(R = 1|Y, X)$ are parametric, there is still an identifiability issue and the use of an instrument satisfying (2)-(3) may be needed. See, for example, Wang et al. (2014) and Miao et al. (2016).

In applications, however, an instrument satisfying (2)-(3) is not given and we must search for an instrument using observed data. The purpose of this paper is to propose and study a method for instrument search from the given covariate vector X, assuming that an instrument satisfying (2)-(3) exists. After an introduction of a pseudo-likelihood method in Sect. 2, we propose a pseudo-likelihood-based maximum ratio criterion to select an instrument. Although our method requires a model such as (1), we can combine our method with model selection regarding (1) in Fang and Shao (2016) to select instrument and model together. In Sect. 3, we establish that, with probability tending to 1 as the sample size $N \to \infty$, while the dimension of X remains fixed, our selected instrument equals an instrument

satisfying (2)-(3). To complement theoretical work, we carry out some simulations in Sect. 4 to examine finite sample properties. For illustration, we apply the proposed method to a real data set in Sect. 5. All technical details are given in an Appendix.

2 Method

We use the notation in Sect. 1, i.e., Y is a response subject to nonresponse, R is the indicator of observing Y, and X is a covariate vector with no missing values. Under assumption (1), our goal is to estimate θ in (1) based on a random sample $\{(y_i, x_i, r_i), i = 1, \ldots, N\}$ from (Y, X, R), where y_i is observed if and only if $r_i = 1$. When asymptotic properties are studied, we consider $N \to \infty$, while the dimension of X remains fixed.

Note that (2) implies that:

$$p(Z|Y, U, R = 1) = p(Z|Y, U) = \frac{p(Y|U, Z)p(U, Z)}{\int p(Y|U, z)p(U, z)dz}.$$

If we know which components of X formed Z satisfying (2)-(3), then we can estimate θ by $\widehat{\theta}$ that maximizes the following pseudo-likelihood based on data with $r_i = 1$:

$$\prod_{i \le N,\ r_i=1} p(z_i|y_i, u_i) = \prod_{i \le N,\ r_i=1} \frac{f(y_i|u_i, z_i;\theta)\widehat{p}(u_i|z_i)d\widehat{F}(z_i)}{\int f(y_i|u_i, z;\theta)\widehat{p}(u_i|z)d\widehat{F}(z)} \tag{4}$$

(Tang et al. 2003; Zhao and Shao 2015), where $\widehat{F}$ is the empirical cumulative distribution function of Z, $\widehat{p}(u|z)$ is a consistent estimator of $p(u|z)$ using observed $x_i = (u_i, z_i)$ and an available method in the literature, either parametrically or nonparametrically, e.g., Lipsitz and Ibrahim (1996), Ibrahim et al. (1999), Zhao and Shao (2015), and Chen et al. (2018). Note that condition (3) ensures that the likelihood function in (4) is a non-constant function of θ. According to Zhao and Shao (2015), as long as there is a non-binary Z satisfying (2)-(3), $\widehat{\theta}$ is consistent and asymptotically normal as $N \to \infty$.

Note that Z satisfying (2) guarantees the validity of pseudo-likelihood (4), whereas Z satisfying (3) ensures that likelihood (4) depends on θ, so that we can estimate θ by maximizing (4).

We now consider how to search for an instrument Z satisfying (2)-(3) from X based on the sample data. For a given candidate Z, a sub-vector of X, our idea is to check whether Z can be an instrument by comparing two estimators of the condition distribution $F_1(z) = P(Z \le z|R = 1)$, where z is a possible value for Z and, for two vectors a and b, $a \le b$ means that all components of a are less than or equal to the corresponding components of b. One estimator is the empirical cumulative distribution function:

$$\widehat{F}_1(z) = \sum_{i=1}^{N} r_i I(z_i \le z) \Big/ \sum_{i=1}^{N} r_i,$$

which does not depend on any model or instrument and, hence, is always consistent, where $I(\cdot)$ is the indicator function. The other estimator is based on the pseudo likelihood (4) and $\widehat{\theta}$. Under condition (2):

$$\begin{aligned} F_1(z) =& E\{I(Z \le z)|R = 1\} \\ =& E[E\{I(Z \le z)|Y, U, R = 1\}|R = 1] \\ =& E[E\{I(Z \le z)|Y, U\}|R = 1] \\ \approx& \sum_{i=1}^{N} r_i E\big\{I(Z \le z)|y_i, u_i\big\} \Big/ \sum_{i=1}^{N} r_i, \end{aligned} \tag{5}$$

where the approximation is valid because of the law of large numbers. Also:

$$\begin{aligned} E\big\{I(Z \le z)|y_i, u_i\big\} =& \int I(t \le z) p(t|y_i, u_i) dt \\ =& \frac{\int I(t \le z) p(y_i|u_i, t) p(u_i|t) dF(t)}{\int p(y_i|u_i, t) p(u_i|t) dF(t)} \\ \approx& \frac{\sum_{j=1}^{N} f(y_i|u_i, z_j; \widehat{\theta}) \widehat{p}(u_i|z_j) I(z_j \le z)}{\sum_{j=1}^{N} f(y_i|u_i, z_j; \widehat{\theta}) \widehat{p}(u_i|z_j)}, \end{aligned} \tag{6}$$

where F is the distribution function of Z and the approximation is valid if (1)-(2) hold and $\widehat{\theta}$ and $\widehat{p}(u|z)$ are consistent. Hence, using (5) and (6), we estimate $F_1(z)$ by:

$$\widetilde{F}_1(z) = \sum_{i=1}^{N} r_i \frac{\sum_{j=1}^{N} f(y_i|u_i, z_j; \widehat{\theta}) \widehat{p}(u_i|z_j) I(z_j \le z)}{\sum_{j=1}^{N} f(y_i|u_i, z_j; \widehat{\theta}) \widehat{p}(u_i|z_j)} \Big/ \sum_{i=1}^{N} r_i. \tag{7}$$

If Z is an instrument, then $\widetilde{F}_1(z)$ in (7) should be close to the empirical cumulative distribution $\widehat{F}_1(z)$; otherwise, the two estimators may not be close to each other.

Following Fang and Shao (2016), a natural idea is to select Z based on the following expected distance between $\widetilde{F}_1(z)$ and $\widehat{F}_1(z)$:

$$\text{VC} = \frac{1}{N} \sum_{i=1}^{N} \big|\widetilde{F}_1(z_i) - \widehat{F}_1(z_i)\big|, \tag{8}$$

where VC means "validation criterion". We may search all possible Z and take the one with the smallest VC value as our estimated instrument $\widehat{Z}$.

However, there are two serious issues. First, to find an instrument from a q-dimensional X, we need to consider all possible $2^q - 1$ non-empty sub-vectors of X as candidates. Even if we do not consider high-dimensional X, a search over $2^q - 1$ candidates is still computationally infeasible when q is not very small. Also, if the dimension of Z is not very small, the estimation of $F_1(z)$ is quite a challenge itself.

The second issue is that validation criterion (8) can check whether the candidate Z satisfies (2), but cannot check whether Z satisfies (3). More precisely, if Z satisfies both (2) and (3), i.e., Z is an instrument, then VC $\to 0$ in probability as $N \to \infty$, because both $\widetilde{F}_1$ and $\widehat{F}_1$ are consistent for F_1; if Z does not satisfy (2), then $\widetilde{F}_1$ does not converge to F_1 and, hence, VC does not converge to 0; however, if Z satisfies (2) but not (3), then approximation (6) is still good because $f(y_i|u_i, z_j;\widehat{\theta}) = f(y_i|u_i;\widehat{\theta})$ can be canceled from the numerator and denominator on the right-hand side of (6) and, thus, VC still converges in probability to 0 even if $\widehat{\theta}$ is not consistent. As discussed after deriving pseudo-likelihood (4), we need a Z satisfying both (2) and (3).

To address these two issues, we proposed a whole new two-step instrument search procedure as follows.

In the first step, we prepare a candidate instrument set including only one- or two-dimensional covariates and calculate their VC values. As discussed in Zhao and Shao (2015), a single binary covariate alone cannot be used as an instrument, since it does not provide enough information to identify all population parameters, except for some special situations. Thus, other than the non-binary covariates, we combine each binary covariate with another single covariate (binary or not) to form a vector of candidates, $(Z_1, ..., Z_p)$, where p may be different from the original dimension of X. For example, if $X = (X_1, ..., X_q)$ is q-dimensional, X_q is binary and all $X_1, ..., X_{q-1}$ are non-binary, then $p = 2q - 2$, $Z_j = X_j$, $j = 1, ..., q-1$ ($q-1$ single covariates), and $Z_j = (Z_{j-q+1}, X_q)$, $j = q, ..., p$ ($q-1$ combined covariates). See Sect. 5 for an example. We assume that:

$$(Z_1, ..., Z_p) \text{ contains at least one } Z_k \text{ satisfying (2) – (3).} \tag{9}$$

There may be other ways to avoid selecting a binary covariate as instrument, e.g., if X_q is the only binary covariate, we may set $(Z_1, ..., Z_p) = (X_1, ..., X_{q-1})$ to exclude X_q, in which case $p = q - 1$.

For each Z_k, $k = 1, ..., p$, let U_k be the sub-vector of X with components not in k. For the kth split of X into (U_k, Z_k), let u_{ki} and z_{ki} be observed U_k and Z_k, respectively, $\widehat{\theta}_k$ be the maximizer of (4) with $u_i = u_{ki}$, $z_i = z_{ki}$, and $\widehat{p}(u_i|z_i) = \widehat{p}(u_{ki}|z_{ki})$, and let $\widetilde{F}_{1k}(z)$ be defined by the right-hand side of (7) with $\widehat{\theta} = \widehat{\theta}_k$, $u_i = u_{ki}$, $z_i = z_{ki}$, and $\widehat{p}(u_i|z_i) = \widehat{p}(u_{ki}|z_{ki})$. To validate whether Z_k is a correct instrument, we calculate:

$$\text{VC}(k) = \frac{1}{N} \sum_{i=1}^{N} \left| \widetilde{F}_{1k}(z_{ki}) - \widehat{F}_1(z_{ki}) \right|. \tag{10}$$

In the second step, we try to find all Z_k satisfying (2) and then put them together as a sub-vector of our searched instrument $\widehat{Z}$. This step is based on the following fact. Let $S = \{1 \le k \le p, Z_k \text{ satisfies (2)}\}$ and Z_S be the sub-vector containing all Z_k's with $k \in S$. Then, Z_S satisfies (2), because each of its component satisfies (2). Also, assumption (9) together with the definition of S implies that there is at least one $k \in S$, such that Z_k satisfies (3); consequently, Z_S including this Z_k also satisfies (3). Hence, Z_S is an instrument satisfying both (2) and (3), although we may not know which Z_k in Z_S satisfying (3).

It remains to find out how we identify the set S. Since a Z_k with a small $\mathrm{VC}(k)$ is likely to be in Z_S, a simple method is to estimate S by $\widehat{S}_\tau = \{k : \mathrm{VC}(k) < \tau\}$, where $\tau > 0$ is a pre-specified threshold. In application, however, it may be difficult to find a τ to split $\{\mathrm{VC}(k), k = 1, \ldots, p\}$ accurately because of the variability in $\mathrm{VC}(k)$. Instead, we propose the following method similar to that in Huang et al. (2014) for feature screening. Let $\{l_1, ..., l_p\}$ be a permutation of $\{1, \ldots, p\}$, such that $\mathrm{VC}(l_1) \le \mathrm{VC}(l_2) \le \cdots \le \mathrm{VC}(l_p)$. Then, our estimator of S is $\widehat{S} = \{l_1, ..., l_{\widehat{d}}\}$, where:

$$\widehat{d} = \begin{cases} \arg\min_{1 \le j \le p-1} \mathrm{VC}(l_j)/\mathrm{VC}(l_{j+1}) & \text{if } \dfrac{\mathrm{VC}(l_p) - \mathrm{VC}(l_1)}{\mathrm{VC}(l_1)} > (\log N)^{1/2}, \\ p & \text{if } \dfrac{\mathrm{VC}(l_p) - \mathrm{VC}(l_1)}{\mathrm{VC}(l_1)} \le (\log N)^{1/2}. \end{cases} \tag{11}$$

This is based on the following facts as $N \to \infty$:

(a) When $k \in S$, $\mathrm{VC}(k) \to 0$ in probability at the rate $N^{-1/2}$. When $k \notin S$, $\mathrm{VC}(k) \to$ a positive quantity in probability.
(b) If the dimension of S is $d \le p - 1$, then $\mathrm{VC}(l_j)/\mathrm{VC}(l_{j+1})$ converges to a positive constant in probability either when $j + 1 \le d$ or when $j \ge d + 1$, but $\mathrm{VC}(l_d)/\mathrm{VC}(l_{d+1}) \to 0$ in probability.
(c) If the dimension of S is p, then $\mathrm{VC}(l_p) - \mathrm{VC}(l_1) \le (\log N)^{1/2}\mathrm{VC}(l_1)$ with probability tending to 1 as $N \to \infty$.

Our selected instrument is then $\widehat{Z} = Z_{\widehat{S}}$ and the corresponding U is $\widehat{U}$ containing components not in $\widehat{Z}$, i.e., $X = (\widehat{U}, \widehat{Z})$.

To end this section, we make the following two remarks.

Remark 1 Although the validation criterion defined in (8) is the same as the VC defined in Fang and Shao (2016), the proposed method is essentially different from the method in Fang and Shao (2016). First, Fang and Shao (2016) focuses on model selection regarding (1) with a given correct instrument Z, while the proposed method focuses on instrument search for a Z. Second, Fang and Shao (2016) considers all $2^q - 1$ non-empty sub-vectors of X as candidates, which is computationally infeasible as long as q is not very small. The proposed method only needs to consider p candidates, where p is typically linear or quadratic in q. Third, Fang and Shao (2016) estimates distribution function of a Z with possibly large dimension, which could be a difficult task, whereas the proposed method only needs to estimate low-dimensional distribution function, since the candidate Z_k in the first step is univariate or bivariate.

Remark 2 The factor $\log N$ in (11) can be replaced by any sequence a_N satisfying $a_N/N^{1/2} \to 0$ as $N \to \infty$. This is a common phenomenon in model/variable selection. We choose $\log N$, because it is used in the well-known BIC and it performs well in our simulation studies.

3 Asymptotic theory

We now establish some asymptotic properties of the validation criterion and $\widehat{Z}$. First, we study the asymptotic properties of $\widehat{\theta}_k$. When Z_k is a correct instrument satisfying (2)-(3), as shown in Zhao and Shao (2015), $\widehat{\theta}_k$ is consistent for θ and asymptotically normal under some regularity conditions. When Z_k does not necessarily satisfy (2), i.e., Z_k may be in the propensity model, we have the following result to show the property of $\widehat{\theta}_k$ under a misspecified instrument.

Theorem 1 *Assume the following regularity conditions.*

(*C*1) *The estimator* $\widehat{\eta}_k = \widehat{p}(u_{ki}|z_{ki})$ *of* $\eta_k = p(u_{ki}|z_{ki})$ *used in* (10) *is constructed using observed covariate data and either a correctly specified parametric model on* η_k *or a nonparametric kernel method, so that* $\widehat{\eta}_k$ *is consistent and asymptotically normal.*

(*C*2) *Let*

$$p(Z_k|Y, U_k;\theta_k, \eta_k, F) = \frac{f(Y|U_k, Z_k;\theta_k)\eta_k(U_k|Z_k)dF(Z_k)}{\int f(Y|U_k, z;\theta_k)\eta_k(U_k|z)dF(z)}$$

and $I_k(\theta_k, \eta_k) = E[R \log p(Z_k|Y, U_k)/p(Z_k|Y, U_k;\theta_k, \eta_k, F)]$, *where* F *is the true distribution of* Z_k. *For* $\widetilde{\eta}_k$ *in a neighborhood of* η_k, $I_k(\theta, \widetilde{\eta}_k)$ *has a unique minimum over* θ.

(*C*3) *Write* $H_k(W;\theta_k, \eta_k, F) = R \log p(Z_k|Y, U_k;\theta_k, \eta_k, F)$, *where* $W=(Y, X, R)$ *and* $w_i = (y_i, x_i, r_i)$. *For some* $\epsilon_1 > 0$ *and* $\epsilon_2 > 0$, *as* $N \to \infty$:

$$\sup_{\substack{||\widetilde{\eta}_k - \eta_k|| < \epsilon_1 \\ ||G - F|| < \epsilon_2}} \left| \frac{1}{N} \sum_{i=1}^{N} H_k(w_i;\theta_k, \widetilde{\eta}_k, G) - E\{H_k(W;\theta_k, \widetilde{\eta}_k, G)\} \right| \to 0,$$

and $E\{H_k(W;\theta_k, \widehat{\eta}_k, \widehat{F}) - H_k(W;\theta_k, \eta_k, F)\} \to 0$ *in probability.*

(*C*4) *Assumption* (1) *holds and* $f(Y|X;\theta)$ *is twice continuously differentiable with respect to* θ *on a bounded and closed set* Θ *in the Euclidean space of a fixed dimension, and the matrix* $E\{-\partial^2 H_k(W;\theta, \eta_{k0}, F)/\partial\theta\partial\theta^T|_{\theta=\theta_k^*}\}$ *is positive definite, where the expectation is with respect to the true distribution of* W *and* θ_k^* *is an interior point of* Θ *that minimizes* $I_k(\theta, \eta_k)$.

Then, $\widehat{\theta}_k$ *is consistent for* θ_k^* *and asymptotically normal as* $N \to \infty$.

Condition (C1) means that a consistent and asymptotically normal estimator $\widehat{\eta}_k$ is required. If a correct parametric model for η_k is not possible, then we have to apply a nonparametric method such as kernel (Zhao and Shao 2015; Chen et al. 2018). If the dimension of X is not low, then dimension reduction needs to be applied (Chen et al. 2018).

Define:

$$F_{1k}(z_k) = E_{Y,U|R=1}\left\{ \frac{\int f(Y|U_k, Z_k; \theta_k^*)\eta_k(U_k|Z_k)I(Z_k \le z_k)dF(Z_k)}{\int f(Y|U_k, Z_k; \theta_k^*)\eta_k(U_k|Z_k)dF(Z_k)} \right\} \quad (12)$$

and

$$\Delta_k = E|F_{1k}(Z) - F_1(Z)|. \quad (13)$$

Since $\widehat{\theta}_k \to \theta_k^*$ and $\widehat{\eta}_k \to \eta_k$ in probability, $\widehat{F}_{1k}(z_k)$ also converges to $F_{1k}(z_k)$ in probability as $N \to \infty$. Then, VC(k) defined in (10) converges in probability to Δ_k. If Z_k is a correct instrument satisfying (2)-(3), then $\Delta_k = 0$. Note that, if Z_k satisfies (2) but not (3), i.e., Z_k could be excluded from both $p(Y|X)$ and $P(R = 1|Y, X)$, condition (C2) in Theorem 1 is not satisfied and θ_k^* is not identifiable. Hence, $\widehat{\theta}_k$ is not a consistent estimator of θ_k^*. Fortunately, in this case, $f(Y|U_k, Z_k; \theta_k^*) = f(Y|U_k; \theta_k^*)$, and it can also be canceled from the numerator and denominator on the right-hand side of (12), which shows that $\Delta_k = 0$ and VC(k) still converges to Δ_k although $\widehat{\theta}_k$ is not consistent when Z_k satisfies (2) but not (3). Hence, if we assume that $\Delta_k > 0$ unless Z_k satisfies (2), then the covariates included in the propensity model could be distinguished and the union of rest covariates could be selected as an instrument.

Theorem 2 *Let* $S = \{1 \le k \le p, Z_k$ satisfies (2)$\}$, Z_S *be the sub-vector containing all* Z_k*'s with* $k \in S$, $\widehat{d}$ *be given by* (11), $\widehat{S}$ *be the index set contains indices of* $\widehat{d}$ *smallest* VC(k)*'s, and* $\widehat{Z} = Z_{\widehat{S}}$ *be the vector containing all* Z_k*'s with* $k \in \widehat{S}$. *Assume* (1), (9), *(C1)–(C4) in Theorem* 1, *and* $\Delta_k > 0$ *unless* Z_k *satisfies* (2). *Then* $P(\widehat{S} = S) \to 1$ *and* $P(\widehat{Z} = Z_S) \to 1$ *as* $N \to \infty$ *while* p *remains fixed; that is, with probability tending to 1,* $\widehat{Z}$ *is the vector of all covariates satisfying* (2). *Furthermore, with probability tending to 1,* $\widehat{Z}$ *satisfies* (3).

Once $Z = \widehat{Z}$ is selected to be an instrument, the rest of components in X forms $\widehat{U}$, which is the most compact U in the propensity $P(R = 1|Y, U)$. After instrument selection, θ in model (1) can be estimated by maximizing the pseudo-likelihood (4) with $Z = \widehat{Z}$ and $U = \widehat{U}$. For inference on θ, we recommend the bootstrap, since the form of asymptotic covariance matrix of $\widehat{\theta}$ is complicated.

Although our proposed method requires a model such as (1), we can combine our method with the model selection method regarding (1) in Fang and Shao (2016) to select instrument and model together. Specifically, we can consider several candidate models $f_l(Y|X, \theta_l)$, $l = 1, ..., L$. For each candidate model l, the instrument search procedure can be carried as described in Sect. 2 by treating $f_l(Y|X, \theta_l)$ as the model. Based on $f_l(Y|X, \theta_l)$ and the selected instrument $\widehat{Z}_l$, we calculate the penalized validation criterion (PVC) in Fang and Shao (2016), which is the same as VC(k) in (10) plus a penalty term that converges to 0 as $N \to \infty$. Assume that Δ_k defined in (13) is positive when model $f_l(Y|X, \theta_l)$ is wrong. Then, the PVC does not converges to 0 when $f_l(Y|X, \theta_l)$ is wrong. Thus, if we select the model with smallest PVC value, then the combined instrument search and model selection procedure finds a correct model and a correct instrument. We illustrate this combined method in Sect. 5.

In theory, Theorems 1–2 ensure that estimation and inference based on selected variables and pseudo-likelihood are asymptotically valid. Although our limited

simulation results show that the finite sample bias of $\widehat{\theta}$ is negligible, there may be "winner's curse" bias in some applications, pointed early by Pötscher (1991) and Zhang (1992) and more recently by Leeb and Pötscher (2006) and Bachoc et al. (2019). One way to mitigate winner's curse bias is to use an additional independent dataset for variable selection, but it is costly. Alternatively, one may split the data set into two parts, one for variable selection and the other for inference afterwards. This deserves further research.

4 Simulation

In this section, we study the finite-sample performance of the proposed method in terms of the rate that we select instrument, and the performance of estimators of θ and $E(Y)$ based on the pseudo-likelihood and the selected instrument. All the results are based on 1000 simulation replications and three sample sizes $N = 100,\ 200,\ 500$.

4.1 Simulation study 1

In the first simulation study, we consider a three-dimensional covariate vector $X = (X_1, X_2, X_3)$ with no redundant covariate in $p(Y|X)$. We consider independent X_j's, $X_j \sim$ chi-square with one degree of freedom, $j = 1, 2, 3$, and $f(Y|X;\theta)$ to be the density of $N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \sigma^2)$, where $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2) = (5, 2, 2, 2, 3)$. For the nonresponse propensity, we consider the following three cases:

1. $P(R=1|Y,X) = \left[1+\exp\{-Y+8\}\right]^{-1}$, $U = \emptyset, Z_S = X$;
2. $P(R=1|Y,X) = \left[1+\exp\left\{-Y(4X_1-1)\right\}\right]^{-1}$, $U = X_1, Z_S = (X_2, X_3)$;
3. $P(R=1|Y,X) = \left[1+\exp\left\{-1.2Y(8X_1X_2-1)\right\}\right]^{-1}$, $U = (X_1, X_2), Z_S = X_3$.

The unconditional response rates for these three cases are 68%, 63% and 52%, respectively. Since all covariates are continuous, the candidate instrument set in the first step is just (X_1, X_2, X_3), but the final selected instrument $\widehat{Z} = Z_{\widehat{S}}$ has $2^3 - 1 = 7$ possible results. To estimate $p(U|Z)$ in the pseudo-likelihood (4), we apply the nonparametric kernel method (Zhao and Shao 2015; Chen et al. 2018).

Table 1 reports the number of times of selecting each possible instrument by the proposed search procedure. It can be seen that the proposed method can select the instrument Z_S with empirical probability much higher than those for other candidates. When the sample size is 200 or larger, the probability of correctly selecting Z_S nearly equals 1.

Besides instrument selection, we also consider the estimation of θ and $E(Y)$ using different instruments and the pseudo-likelihood (4). Table 2 reports the empirical means and standard derivations of estimators when $N = 200$. With missing data, the only data-adaptive estimators are those based on $\widehat{Z}$, and the rest estimators based on a fixed choice of instrument are entered for comparison. In Case 1, the vector of all instruments is the whole covariate vector, i.e., $Z_S = (X_1, X_2, X_3)$; but any non-empty sub-vector of X is also a correct instrument satisfying (2)-(3). In Case 2, the vector

Table 1 Number of times validation criterion (11) selects each instrument in 1000 simulations

Case	Z_S	N	Selected instrument by (11)						
			X_1	X_2	X_3	(X_1, X_2)	(X_1, X_3)	(X_2, X_3)	(X_1, X_2, X_3)
1	(X_1, X_2, X_3)	100	7	7	9	8	6	4	959
		200	4	7	6	5	3	4	971
		500	3	2	2	3	3	4	983
2	(X_2, X_3)	100	0	15	19	0	0	965	1
		200	0	10	8	0	0	982	0
		500	0	0	0	0	0	1000	0
3	X_3	100	2	3	701	0	53	57	184
		200	0	0	939	0	11	9	41
		500	0	0	1000	0	0	0	0

of all instruments is $Z_S = (X_2, X_3)$; and either X_2 or X_3 is a correct one-dimensional instrument. In Case 3, only $Z_S = X_3$ is the correct instrument. Table 2 shows the results when a correct instrument or our proposed $\widehat{Z}$ is used. It can be seen that estimators are almost unbiased, but the use of true Z_S produces the most efficient estimators, and estimators based on the proposed $\widehat{Z}$ are almost the same as those based on Z_S. On the other hand, when we use a wrong instrument (for example, in Case 2, X_1, (X_1, X_2), (X_1, X_3), and (X_1, X_2, X_3) are wrong instruments), some estimators are seriously biased. The last row for each case of Table 2 reports the results for estimators when there are no missing data, which provides a standard of the best we can do.

4.2 Simulation study 2

In the second simulation study, we consider the situation where $p(Y|X)$ involves some redundant covariates and $X = (X_1, \ldots, X_{10})$ is 10-dimensional. Here, Y and (X_1, X_2, X_3) are generated the same as those in simulation study 1 in Sect. 4.1, and $X_4, \ldots, X_{10}$ are generated independently from the standard normal distribution and are redundant, i.e., $p(Y|X) = p(X_1, X_2, X_3)$. In this experiment, we consider the following three models for (1):

$$
\begin{aligned}
M_1 &: Y|X \sim N\big(\beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10},\, \sigma^2\big),\\
M_2 &: Y|X \sim N\big(\beta_0 + \gamma_1 X_1^2 + \cdots + \gamma_{10} X_{10}^2,\, \sigma^2\big),\\
M_3 &: Y|X \sim N\big(\beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10} + \gamma_1 X_1^2 + \cdots + \gamma_{10} X_{10}^2,\, \sigma^2\big),
\end{aligned}
$$

where M_1 is a correct model, M_2 is a wrong model, and M_3 is correct but overfitted. The best model, i.e., the most compact correct model, is M_1 with $\beta_4 = \cdots = \beta_{10} = 0$. Therefore, in this experiment, besides instrument search, we also consider model selection as discussed in Remark 2 in the end of Sect. 2.

Table 2 Mean and standard deviation (in parentheses) of estimators using different Zs when $N = 200$ based on 1000 simulations

Case	Z used	Parameter					
		$\beta_0 = 5$	$\beta_1 = 2$	$\beta_2 = 2$	$\beta_3 = 2$	$\sigma^2 = 3$	$E(Y) = 11$
1	X_1	5.00 (0.39)	2.00 (0.13)	2.01 (0.14)	2.01 (0.14)	2.91 (0.49)	11.01 (0.40)
	X_2	4.99 (0.40)	2.01 (0.15)	2.01 (0.12)	2.01 (0.15)	2.93 (0.49)	11.00 (0.40)
	X_3	5.01 (0.38)	2.00 (0.14)	2.01 (0.15)	2.00 (0.12)	2.91 (0.49)	11.01 (0.41)
	(X_1, X_2)	5.00 (0.33)	2.00 (0.11)	2.00 (0.11)	2.00 (0.12)	2.93 (0.43)	10.99 (0.38)
	(X_1, X_3)	5.01 (0.33)	2.00 (0.11)	2.01 (0.12)	2.00 (0.11)	2.91 (0.43)	11.00 (0.38)
	(X_2, X_3)	5.01 (0.32)	2.00 (0.11)	2.00 (0.11)	2.00 (0.11)	2.92 (0.43)	10.99 (0.39)
	(X_1, X_2, X_3)	5.01 (0.30)	2.00 (0.10)	2.00 (0.10)	2.00 (0.10)	2.91 (0.40)	10.99 (0.37)
	$\hat{Z}$	5.01 (0.30)	2.00 (0.10)	2.00 (0.10)	2.00 (0.10)	2.91 (0.40)	10.99 (0.37)
	No missing	4.99 (0.19)	2.00 (0.09)	2.00 (0.09)	2.00 (0.09)	2.93 (0.30)	10.98 (0.36)
2	X_1	3.36 (0.40)	2.48 (0.18)	2.00 (0.14)	2.00 (0.14)	3.84 (0.65)	9.84 (0.50)
	X_2	4.99 (0.44)	2.01 (0.16)	2.00 (0.16)	2.02 (0.19)	2.90 (0.54)	11.01 (0.42)
	X_3	5.00 (0.43)	2.00 (0.16)	2.01 (0.18)	2.01 (0.16)	2.91 (0.53)	11.01 (0.42)
	(X_1, X_2)	4.10 (0.34)	2.23 (0.13)	2.13 (0.14)	2.00 (0.14)	3.29 (0.48)	10.44 (0.44)
	(X_1, X_3)	4.10 (0.34)	2.23 (0.13)	2.00 (0.13)	2.13 (0.14)	3.29 (0.48)	10.44 (0.44)
	(X_2, X_3)	5.00 (0.35)	2.00 (0.13)	2.00 (0.13)	2.00 (0.13)	2.91 (0.44)	10.99 (0.39)
	(X_1, X_2, X_3)	4.43 (0.34)	2.15 (0.14)	2.07 (0.13)	2.07 (0.13)	3.13 (0.43)	10.68 (0.42)
	$\hat{Z}$	5.00 (0.35)	2.00 (0.13)	2.00 (0.14)	2.00 (0.13)	2.91 (0.44)	10.99 (0.39)
	No missing	4.99 (0.19)	2.00 (0.09)	2.00 (0.09)	2.00 (0.09)	2.93 (0.30)	10.98 (0.36)
3	X_1	3.39 (0.49)	2.39 (0.17)	2.20 (0.17)	2.00 (0.17)	3.59 (0.65)	9.95 (0.50)
	X_2	3.37 (0.53)	2.20 (0.18)	2.40 (0.17)	2.00 (0.18)	3.60 (0.68)	9.94 (0.51)
	X_3	4.99 (0.53)	2.00 (0.17)	2.02 (0.18)	2.01 (0.18)	2.90 (0.59)	11.00 (0.43)
	(X_1, X_2)	3.46 (0.42)	2.31 (0.14)	2.31 (0.14)	1.99 (0.16)	3.43 (0.57)	10.05 (0.47)
	(X_1, X_3)	4.11 (0.42)	2.18 (0.14)	2.12 (0.14)	2.09 (0.16)	3.18 (0.51)	10.48 (0.44)
	(X_2, X_3)	4.10 (0.41)	2.11 (0.14)	2.19 (0.13)	2.09 (0.16)	3.19 (0.52)	10.47 (0.44)
	(X_1, X_2, X_3)	3.89 (0.38)	2.21 (0.13)	2.21 (0.13)	2.08 (0.16)	3.22 (0.50)	10.36 (0.45)
	$\hat{Z}$	4.96 (0.57)	2.01 (0.17)	2.02 (0.18)	2.02 (0.18)	2.91 (0.59)	10.98 (0.45)
	No missing	4.99 (0.19)	2.00 (0.09)	2.00 (0.09)	2.00 (0.09)	2.93 (0.30)	10.98 (0.36)

For the propensity, we consider the three cases in Sect. 4.1, and an additional case,

4. $P(R = 1|Y, X) = \left[1 + \exp\left\{-Y(X_4 + 0.5)\right\}\right]^{-1}$, $U = X_4, Z = (X_1, X_2, X_3)$.

The unconditional response rate for Case 4 is 69%. The reason which we consider the additional Case 4 is that it is the case where $p(Y|X)$ is a function of $X_* = (X_1, X_2, X_3)$ and $U = X_4$ is not in X_*, whereas in all Cases 1-3, $U \subset X_* = (X_1, X_2, X_3)$.

Table 3 reports the frequencies in 1000 simulations of correctly selecting the combination of (U, Z_S) and M_1 using the proposed procedure in Sect. 2. It can be seen that, in Cases 1–2 and 4, all correct selection probabilities are higher than 90% and some are

Table 3 Frequency of correctly selecting the combination of (U, Z_S) and M_1 in 1000 replications at each step in simulation study 2

Case	U	Z	N	Step 1	Step 2
1	$\varnothing$	(X_1, X_2, X_3)	100	913	901
			200	928	928
			500	939	939
2	X_1	(X_2, X_3)	100	959	864
			200	982	943
			500	1000	1000
3	(X_1, X_2)	X_3	100	392	340
			200	868	807
			500	998	990
4	X_4	(X_1, X_2, X_3)	100	832	758
			200	985	962
			500	999	999

Step 1: correctly selecting the combination of (U, Z_S) under model M_1

Step 2: correctly selecting the combination of (U, Z_S) and M_1

close to 1 when $N = 200$ and 500, and are between 70% and 90% when $N = 100$. Case 3 is the most difficult situation for instrument search, since Z_* is one-dimensional and X is 10-dimensional; the correct selection probabilities are too low when $N = 100$, but are adequate when $N = 200$, and are close to 1 when $N = 500$.

5 Real data example

To illustrate our proposed instrument selection method, we consider a real data set from the National Health and Nutrition Examination Survey (NHANES) conducted in 2005 by the United States Centers for Disease Control and Prevention, which was also analyzed in Fang and Shao (2016) from model selection perspective. In this data set, Y is the body fat percentage, which is measured by dualenergy X-ray absorptiometry and denoted as dxa. The covariates are age, gender, and body mass index (bmi), i.e., $X = (\text{bmi}, \text{age}, \text{gender})$. As in Fang and Shao (2016), we consider $N = 1591$ middle-aged and elderly people, from whom 393 (24.7%) have missing dxa.

We combine our instrument search method and the model selection method in Fang and Shao (2016) to select instrument and model (1) together as discussed in Remark 2 in the end of Sect. 2. Following Fang and Shao (2016), we consider the following four candidate models:

$$M_1 : Y|X \sim N(\beta_0 + \beta_1 \text{bmi} + \beta_2 \text{age} + \beta_3 \text{gender}, \sigma^2);$$
$$M_2 : Y|X \sim N(\beta_0 + \beta_1 \text{bmi} + \beta_2 \text{age} + \beta_3 \text{gender} + \beta_4 \text{age} \times \text{gender}, \sigma^2);$$
$$M_3 : \log Y|X \sim N(\beta_0 + \beta_1 \log(\text{bmi}) + \beta_2 \text{age} + \beta_3 \text{gender}, \sigma^2);$$
$$M_4 : \log Y|X \sim N(\beta_0 + \beta_1 \log(\text{bmi}) + \beta_2 \text{age} + \beta_3 \text{gender} + \beta_4 \text{age} \times \text{gender}, \sigma^2)$$

Since gender is a binary covariate, as we discussed in Sect. 2, we need to combine gender with either age or bmi to prepare the candidate instrument set in the first step. Then, we have the following $p = 4$ candidates of instrument: $Z_1 =$ age, $Z_2 =$ bmi, $Z_3 =$ (age, gender), and $Z_4 =$ (bmi, gender). Noted that the final selected $\widehat{Z}$ could be the union of these candidates. In this dataset, bmi, age, and gender are almost independent (see also Fang and Shao (2016)), so that the estimation of $p(U|Z)$ has little effect.

For each candidate model $M_j, j = 1, 2, 3, 4$, the proposed instrument search procedure is applied. Table 4 reports the VC(k) value in (10) for each candidate instrument and each model M_j. Based on (11), $\widehat{Z} =$ age is selected as instrument under models M_1 and M_2, and $\widehat{Z} =$ (age, gender) under models M_3 and M_4. Then, we obtain the PVC value defined in (12) of Fang and Shao (2016) using the selected $\widehat{Z}$ under each model (Table 3). The smallest PVC value corresponds to model M_1. As we discussed in Sect. 3, the PVC values corresponding to wrong models do not converge to 0. Thus, M_1 is selected with instrument $\widehat{Z} =$ age and $\widehat{U} =$ (bmi, gender). Using pseudo-likelihood in (4) with the selected model M_1 and instrument $\widehat{Z} =$ age, the estimated parameters are given in the last part of Table 4.

Acknowledgements We thank an associate editor and two referees for their helpful comments. Jun Shao's research was partially supported by the National Scientific Foundation of China (11831008) and the U.S. National Science Foundation grants DMS-1612873 and DMS-1914411. Fang Fang's research was partially supported by the National Scientific Foundation of China (11831008, 11601156, and 11771146).

Appendix: Proofs

Proof of Theorem 1: As $f(Y|X;\theta)$ is continuous with respect to θ on a bounded and closed set Θ, for a given $\widetilde{\eta}_k$, $I_k(\theta_k, \widetilde{\eta}_k)$ must attain a minimum in Θ. Condition C2, which is similar to Assumption A3 in White (1982), guarantees that there exists a

Table 4 Instrument and model selection and estimation results in NHANES dataset

	Model			
	M_1	M_2	M_3	M_4
VC(1) with $Z_1 =$ age	0.0020	0.0020	0.0020	0.0024
VC(2) with $Z_2 =$ bmi	0.0068	0.0067	0.0100	0.0100
VC(3) with $Z_3 =$ (age, gender)	0.0081	0.0073	0.0025	0.0024
VC(4) with $Z_4 =$ (bmi, gender)	0.0180	0.0178	0.0094	0.0094
Selected $\widehat{Z}$ using (11)	age	age	(age, gender)	(age, gender)
PVC value using $\widehat{Z}$ as instrument	0.0065	0.0071	0.0070	0.0074
Final model fitting after M_1 and $\widehat{Z} =$ age are selected				
Parameter	β_0: intercept	β_1: bmi	β_2: age	β_3: gender
Estimate (SE)	7.61 (4.90)	0.75 (0.09)	−11.62 (0.87)	0.21 (0.08)

VC(k) are given by (10) in Sect. 2

PVC are given by (12) in Fang and Shao (2016)

unique maximizer of $E[R\log p(Z_k|Y,U_k;\theta_k,\widetilde{\eta}_k,F)]$ when $\widetilde{\eta}_k$ is in a neighborhood of η_k. Therefore, θ_k^* is identifiable under a misspecified instrument. Under condition C3, following the proof of Theorem 2 in Zhao and Shao (2015), $\widehat{\theta}_k$ converges to θ_k^* in probability as $N\to\infty$.

For each combination of $X=(U_k,Z_k)$, maximizing (4) is the same as maximizing $l_k(\theta_k,\widehat{\eta}_k,\widehat{F})$, where:

$$\begin{aligned}&l_k(\theta_k,\eta_k,F)\\ =&\frac{1}{N}\sum_{i=1}^{N} r_i\{\log p(y_i|u_{ki},z_{ki};\theta_k)-\log\int p(y_i|u_{ki},Z_k;\theta_k)p(u_{ki}|Z_k;\eta_k)dF(Z_k)\}.\end{aligned}$$

Under condition C4, the asymptotic normality of $\widehat{\theta}_k$ follows from Taylor's expansion, $\partial l_k(\theta,\widehat{\eta}_k,\widehat{F})/\partial\theta|_{\theta=\widehat{\theta}_k}=0$, the theory of V-statistics, and the proof of Theorem 3 in Zhao and Shao (2015). □

Proof of Theorem 2: The result of Theorem 2 follows from statements (a)–(c) after formula (11).

First, we prove (a) after formula (11). When Z_k is a correct instrument satisfying (2)-(3), we consider $\mathrm{VC}(k)=\int|\widetilde{F}_{1k}(z_k)-\widehat{F}_1(z_k)|d\widehat{F}(z_k)$. Write $r=\sum_{i=1}^{N} r_i$, and:

$$g_{k0}=\frac{1}{N}\sum_{i=1}^{N} r_i I(Z_k\le z_k),$$

$$g_k(\theta_k,\eta_k,F)=\frac{1}{N}\sum_{i=1}^{N} r_i\frac{\int f(Y|U_k,Z_k;\theta_k)\eta_k(U_k|Z_k)I(Z_k\le z_k)dF(Z_k)}{\int f(Y|U_k,Z_k;\theta_k)\eta_k(U_k|Z_k)dF(Z_k)}.$$

By (6) and central limit theorem, $g_{k0}=E[RI(Z_k\le z_k)]+o_p(N^{-1/2})$ and $g(\theta_k,F)=E[RE\{I(Z_k\le z_k)|Y,U\}]+o_p(N^{-1/2})$. Note that, by the arguments of (5), $E[RI(Z_k\le z_k)]=E[RE\{I(Z_k\le z_k)|Y,U\}]$. Therefore, $g_k(\theta_k,\eta_k,F)=g_{k0}+o_p(N^{-1/2})$. Meanwhile, $\widetilde{F}_{1k}(z_k)=Ng_k(\widehat{\theta}_k,\widehat{\eta}_k,\widehat{F})/r$, $\widehat{F}_1(z_k)=Ng_{k0}/r$ and $N^{1/2}\{\widetilde{F}_{1k}(z_k)-\widehat{F}_1(z_k)\}=N\underset{\sim}{Q}_{Nk}(z_k)$, where $Q_{Nk}(z_k)=N^{1/2}\{g_k(\widehat{\theta}_k,\widehat{\eta}_k,\widehat{F})-g_{k0}\}$. Theorem 1 shows that $\widehat{\theta}_k$ and $\widehat{\eta}_k$ are both consistent and asymptotically normal. Therefore, following the proof of Theorem 2 in Fang and Shao (2016), $Q_{Nk}(z_k)$ converges weakly to a zero-mean Gaussian process. Since $r/N=\sum_{i=1}^{N} r_i/N$ converges to $P(R=1)$ almost surely, $N^{-1/2}\int|\widetilde{F}_{1k}(z_k)-\widehat{F}_1(z_k)|d\widehat{F}(z_k)=\int|Q_{Nk}(z_k)|d\widehat{F}(z_k)/(r/N)$ converges in distribution and is $O_p(1)$. Therefore, $\mathrm{VC}(k)\to 0$ in probability at the rate $N^{-1/2}$ when $k\in S$. When $k\notin S$, $\mathrm{VC}(k)\to\Delta_k$, which is assumed to be positive.

Next, if $d\le p-1$, then $\mathrm{VC}(l_p)\to\Delta_{l_p}>0$ and $\mathrm{VC}(l_1)\to 0$ in probability as $N\to\infty$. Since $(\log N)^{1/2}\mathrm{VC}(l_1)\to 0$ in probability as $N\to\infty$, $\mathrm{VC}(l_p)-\mathrm{VC}(l_1)>(\log N)^{1/2}\mathrm{VC}(l_1)$ with probability tending to 1. Hence, with probability tending to 1, $\widehat{d}=\arg\min_{1\le j\le p-1}\mathrm{VC}(l_j)/\mathrm{VC}(l_{j+1})$. By the proved (a) after formula (11), $\mathrm{VC}(l_j)/\mathrm{VC}(l_{j+1})$ converges to a positive constant in probability when $j\ge d+1$ and $\mathrm{VC}(l_d)/\mathrm{VC}(l_{d+1})\to 0$ in probability. Following the fact that $N^{-\frac{1}{2}}\int|\widetilde{F}_{1k}(z_k)-\widehat{F}_1(z_k)|d\widehat{F}(z_k)=\int|Q_{Nk}(z_k)|d\widehat{F}(z_k)/(r/N)$ converges in distribution

when $k \in S$ and is not degenerate, $\mathrm{VC}(l_j)/\mathrm{VC}(l_{j+1})$ converges to a positive constant in probability when $j+1 \le d$. This proves (b) after formula (11).

If $d = p$, then $Z_S = X$. Hence, $\mathrm{VC}(k) \to 0$ in probability at the rate $N^{-1/2}$ as $N \to \infty$ for $k = 1, \ldots, p$. Since $(\log N)^{1/2}\mathrm{VC}(l_1) \to 0$ at a rate slower than $N^{-1/2}$, $(\log N)^{1/2}\mathrm{VC}(l_1) > \mathrm{VC}(l_p) - \mathrm{VC}(l_1)$ in probability as $N \to \infty$. This proves (c) after formula (11) and the fact that $P(\widehat{d} = p) \to 1$. □

References

Bachoc, F., Leeb, H., Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics*, *47*(3), 1475–1504.

Baker, S. G., Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, *83*(401), 62–69.

Chen, J., Xie, B., Shao, J. (2018). Pseudo likelihood and dimension reduction for data with nonignorable nonresponse. *Statistical Theory and Related Fields*, *2*(2), 196–205.

Fang, F., Shao, J. (2016). Model selection with nonignorable nonresponse. *Biometrika*, *103*(4), 861–874.

Greenlees, J. S., Reece, W. S., Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, *77*(378), 251–261.

Huang, D., Li, R., Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business and Economic Statistics*, *32*(2), 237–244.

Ibrahim, J. G., Lipsitz, S. R., Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *61*(1), 173–190.

Leeb, H., Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, *34*(5), 2554–2591.

Lipsitz, S. R., Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, *83*(4), 916–922.

Little, R. J., Rubin, D. B. (2002). *Statistical analysis with missing data* 2nd ed. New York: Wiley.

Miao, W., Ding, P., Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, *111*(516), 1673–1683.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, *7*(2), 163–185.

Qin, J., Leung, D., Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, *97*(457), 193–200.

Robins, J., Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, *16*(3), 285–319.

Tang, G., Little, R. J., Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*(4), 747–764.

Wang, S., Shao, J., Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*(3), 1097–1116.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, *79*(4), 741–746.

Zhao, J., Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*(512), 1577–1590.